Data Glacier Intern Project Report

Project: Bank Marketing (Campaign) for Farmers Cooperative Bank

Problem Description

Farmers Cooperative bank aims to launch a new term deposit scheme and wants to sell this product to customers. Prior to the launch, the bank plans to start a marketing campaign for the product through various marketing channels like Telephone, SMS, Emails, etc. To save time and to minimize the costs associated with this process, the bank wants to shortlist all the potential customers who have a greater possibility of buying the term deposit product.

This will help the marketing team to start a campaign on a set lot of customers without wasting their resources on any unlikely buyers. To achieve this outcome, we will need to develop a classification model with high accuracy to determine if a customer will subscribe to the term deposit or not based on the available marketing data.

Business Understanding

A new ML model will be developed and deployed on the cloud server subjected to a rigorous evaluation process for selecting the best model to produce optimal results. Bank executives can pass the customer information such as age, income, education, marital status, etc., to predict if the customer would subscribe to the term deposit. The ML application returns the prediction as ‘Yes’ or ‘No’. The team can then consider sending marketing communication to the potential clients based on the prediction made by the ML algorithm.

Data Understanding

The data to be used in the project contains 21 columns and 41188 rows. The data is enclosed in a csv file delimited by semicolon. Description of each column is given below.

Column	Description
Age	Age of the customer
Job	Type of job taken by the customer
Martial	Martial status of the customer
Education	Educational qualification of the customer
Default	Does the customer have a defaulted credit
Housing	Does the customer have a housing loan
Loan	Does the customer have a personal loan
Contact	Communication type for the customer
Month	Last contact month of the customer
Day_of_week	Last contact day of the week
Duration	Last contact duration of the customer
Campaign	Number of times the customer is contacted
Pdays	Number of days passed by after client was contacted
Previous	Number of contacts made to client before campaign
Poutcome	Outcome of the previous campaign for the client
Emp.var.rate	Employment Variation Rate - Quarterly
Cons.price.idx	Consumer price index - Monthly
Cons.conf.idx	Consumer Confidence index - Monthly
Euribor3m	Euribor three-month rate - Daily
Nr.employed	Number of employees - Quarterly
Y	Target Variable – If client subscribed to the plan

Job, martial, education, default, housing, loan, contact, month, and day_of_week and poutcome are categorical variables and the rest of the columns are numeric.

Problems

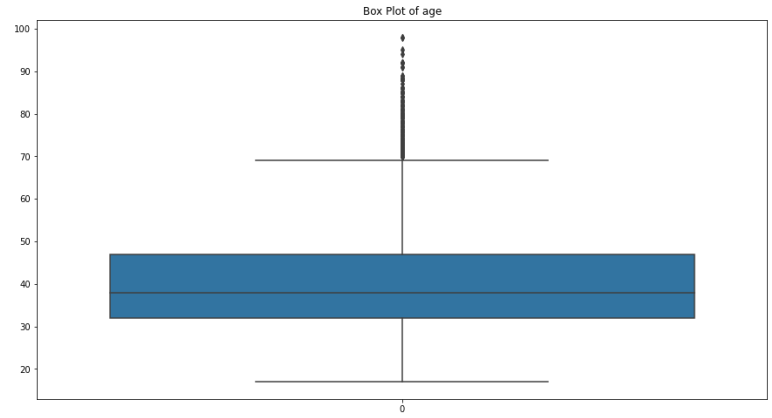
1. **NA Values:** The data implicitly does not contain any None or NaN values. However, some columns contain 'unknown' in some of the rows. We can consider 'unknown' synonymous to NaN since both convey the same meaning. The following images describe the number of unknowns in each column.

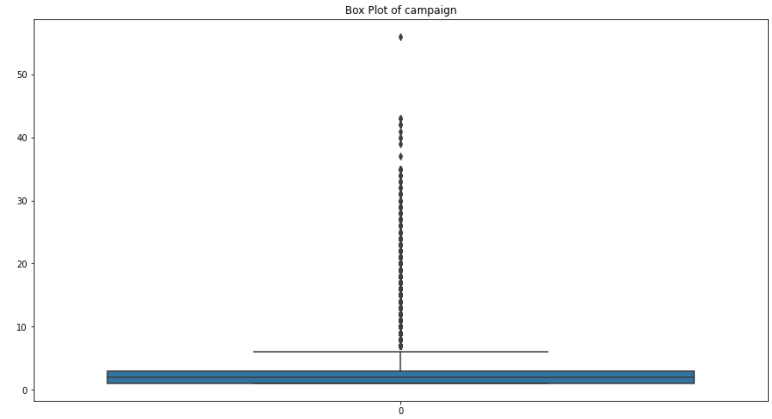
age	0
job	330
marital	80
education	1731
default	8597
housing	990
loan	990
contact	0
month	0
day_of_week	0
duration	0
campaign	0
pdays	0
previous	0
poutcome	0
emp.var.rate	0
cons.price.idx	0
cons.conf.idx	0
euribor3m	0
nr.employed	0
y	0

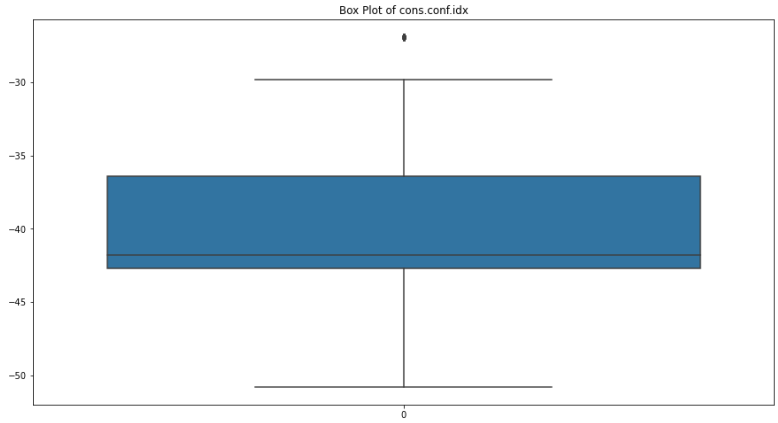
2. **Categorical Variables:** The data contains categorical columns such as job, martial, education, etc. These variables needs to encoded to pass it through a machine learning model. The categorical variables are given below.

job	object
marital	object
education	object
default	object
housing	object
loan	object
contact	object
month	object
day_of_week	object
poutcome	object
y	object

3. **Outliers:** We have created box plots to identify any outliers present in the numeric columns. It is observed that no other columns except age, campaign and cons.conf.idx contain outliers. We can interpret from the below figures that age above 70, campaign above 5 days and consumer confidence index above -30 are all outliers as observed from the below figures.





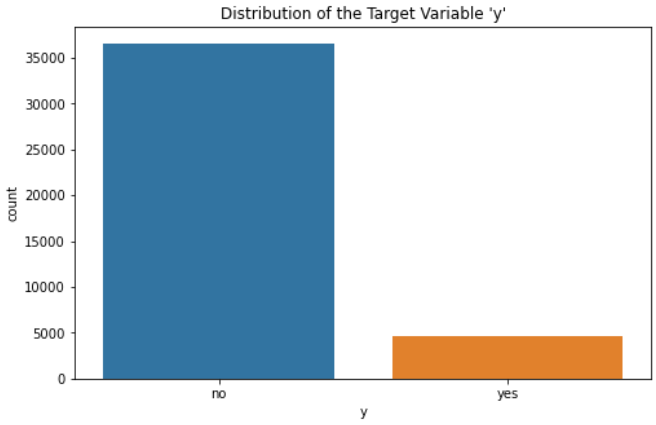


4. **Skewness:** We have observed skewness present in our dataset. The amount of skewness present in each numeric column is given below.

```
age              0.784697
campaign         4.762507
pdays           -4.922190
previous         3.832042
emp.var.rate    -0.724096
cons.price.idx  -0.230888
cons.conf.idx    0.303180
euribor3m       -0.709188
nr.employed     -1.044262
```

5. **Duplicates:** The data contains some duplicate values which might affect the model prediction and must be handled effectively.

6. **Target Variable Distribution:** In our analysis, we found that the more 'No' than 'Yes' in the distribution of the target variable y. This can lead to bias and would negatively impact the performance of the model.



Approaches

- For NA or unknown values, we can impute the unknowns with mean, median or mode value of the column. In this way, we will make the unknown value as close as possible to its true value. It is also possible to replace the unknown with a random sample taken from the data. One more technique is to use a model-based approach to fill the

unknowns by considering information from other columns to predict the unknown value. In this project we shall evaluate the results from all the above-mentioned techniques and choose the method that produces the best results.

- Duplicates contribute to inconsistencies in the prediction. Duplicates also lead to overfitting of the model. In our analysis, we find that there are only 12 duplicate records in the data and thus it is better to drop them before applying any machine learning model.
- Outliers can be handled by truncating them with some upper or lower threshold values. We can also delete the values or apply any mathematical transformations like log, square root, etc., to reduce the impact. We aim to replace the outliers with upper threshold, verify the results and then proceed to other methods based on the results. In this way we can reduce the complexity without inducing more inconsistencies in the prediction.
- Categorical variables need to be encoded to numeric values to pass it to the model. We will be using label encoder or one hot encoder to achieve this. This will create new columns after converting the categorical values to numeric which can then be interpreted by the model.
- For handling the skewness in the data, we aim to normalize the data manually or apply transformations such as logarithmic or box cox. Normalization will convert the data into a normal distribution to have a constant mean and

standard deviation. This will ensure that the prediction is not biased by the skewness present in the data. Logarithmic or box cox transformations will compress the extreme values to smooth the data and introduce normality in the distribution.

- For the Imbalanced target variable distribution, we need to apply under sampling or oversampling techniques to balance the class distribution. Oversampling will increase the number of instances of the minority class and under sampling will reduce the instances of the majority class. Since the dataset is not extremely large, will shall apply oversampling, verify the results and also try out under sampling if required.

Data Cleaning

Duplicates: The data contains 12 duplicates and are removed.

Handling NA values: We have used three imputation methods to handle the NA values.

In method 1, we have replaced the NA values with the most frequent value in the column i.e., mode.

```
NA count before Imputation:
job          330
marital       80
education   1730
default     8596
housing      990
loan         990
dtype: int64

NA count after mode Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

In method 2, we have imputed the NA values with random values taken from the column.

```
NA count before Imputation:
job          330
marital       80
education   1730
default     8596
housing      990
loan         990
dtype: int64

NA count after Random Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

In method 3, we have imputed the NA values by values predicted by a random forest classifier. For this, we have considered the column with NA values as target and the rest of the columns as feature variables.

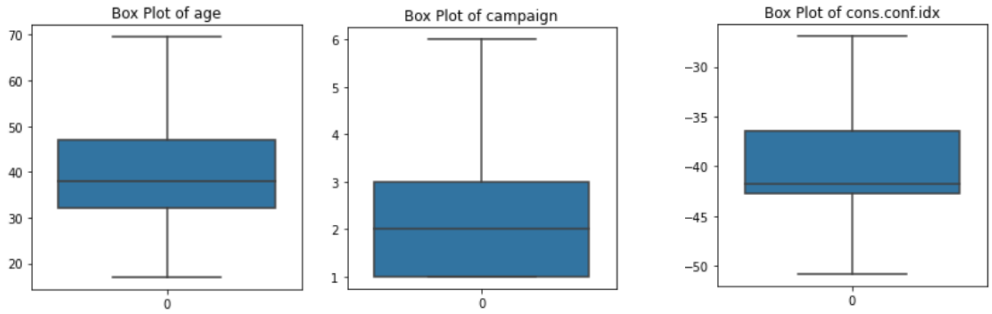
```
NA count before Imputation:
job          330
marital       80
education   1730
default     8596
housing      990
loan         990
dtype: int64

NA count after Model based Imputation:
job          0
marital      0
education    0
default      0
housing      0
loan         0
dtype: int64
```

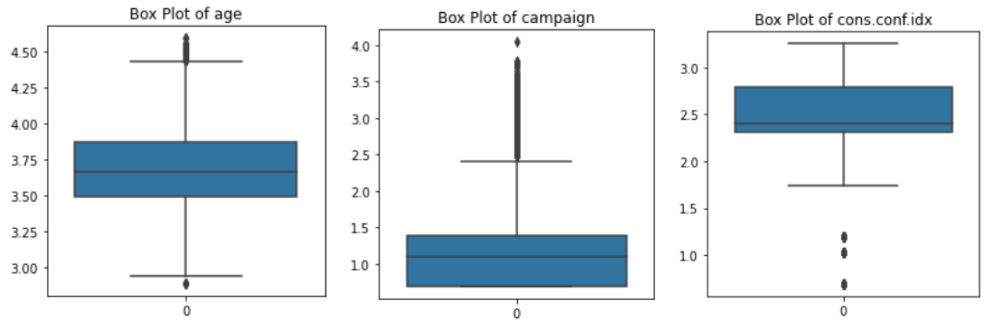
Outlier and Skewness

Previously, we have identified that columns ‘age’, ‘campaign’ and ‘cons.conf.idx’ contain outliers.

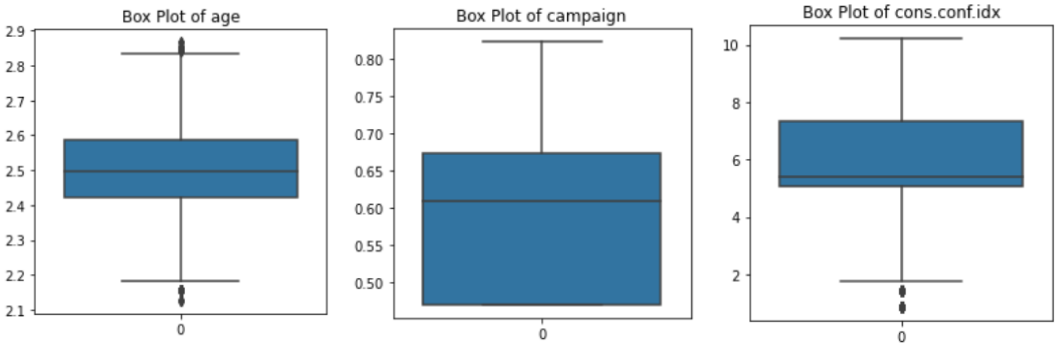
In method 1, we have trimmed the values to lie between upper and lower quartile. In this way we have eliminated almost all the outliers from our data. This method is easy to implement but does not handle skewness in the data.



In method 2, we have applied log transformation to our columns containing outliers. The log transformation does not fully eliminate outliers but eliminates skewness to some extent.



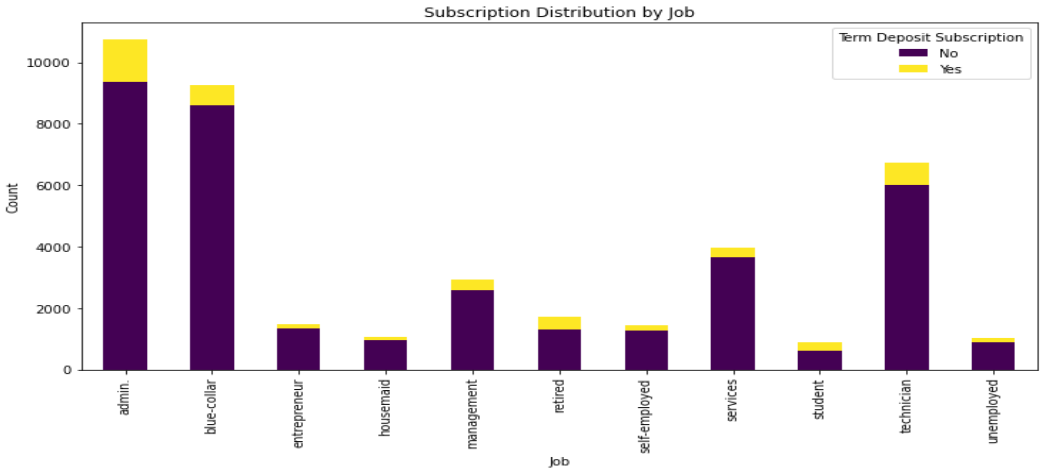
In method 3, we have applied boxcox transformation to our data. Box cox handled most of the outliers while making the data more symmetric and eliminating skewness.



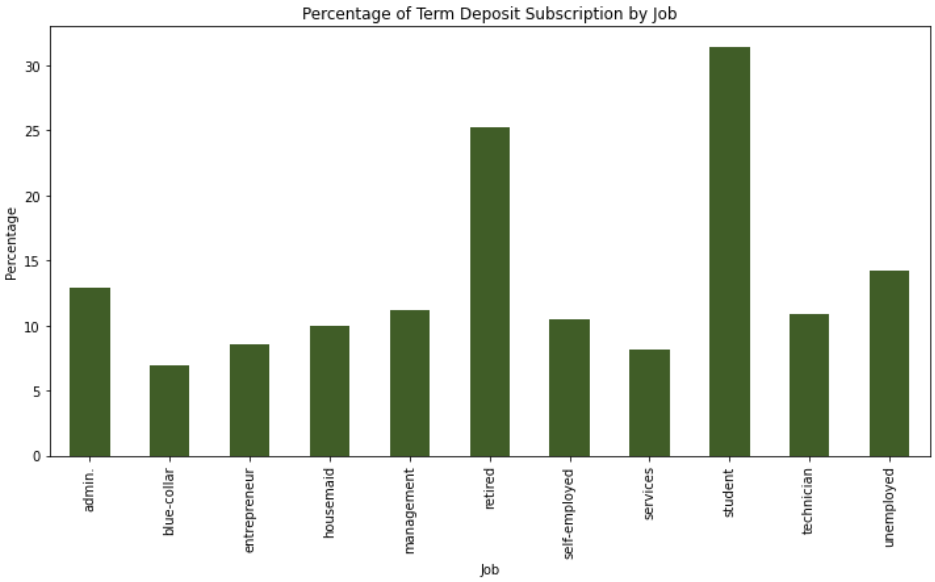
EDA

Most of the customers part of the campaign are into administration and are followed by blue collar workers and technicians. The ratio of those who chose the plan to those who did not choose seems to be

nearly the same for all employment categories. Most of the housemaids, entrepreneurs and umemployed customers did not choose the plan

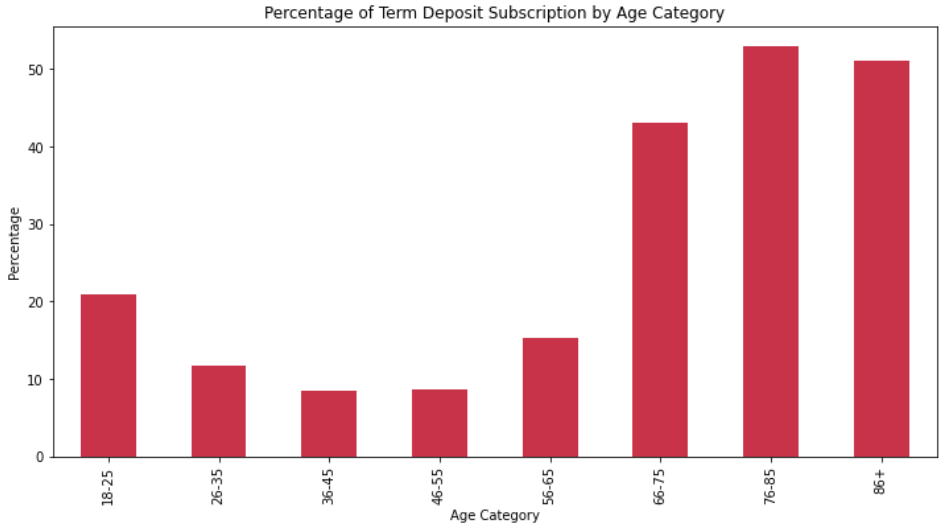


It is also observed that more than 30 percent of the student customers have subscribed to the plan and the bank might need to focus more on the students and retired individuals to sell the term deposit.

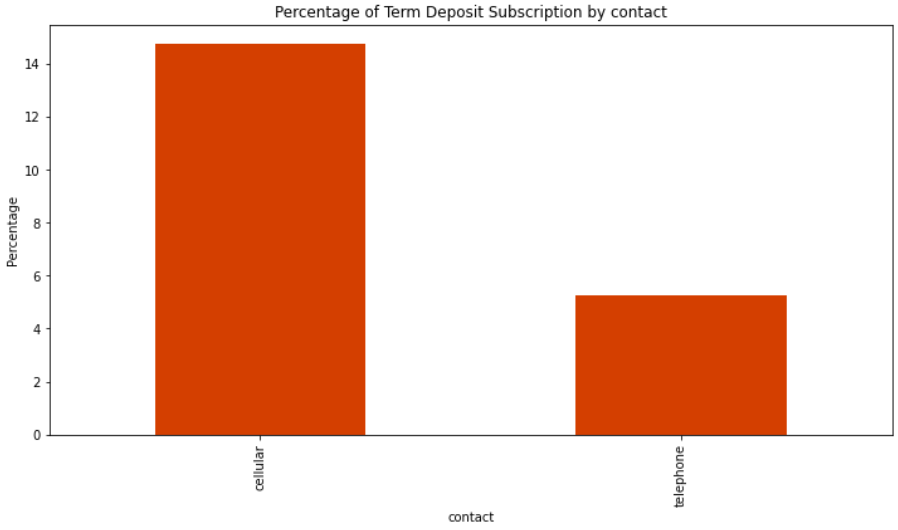


Age is also contributing to the subscription. Young customers seem to be disinterested in the plan. 20 percent of the clients aged 18-25 subscribed to the plan and the results seem poor for middle aged group. 40-50

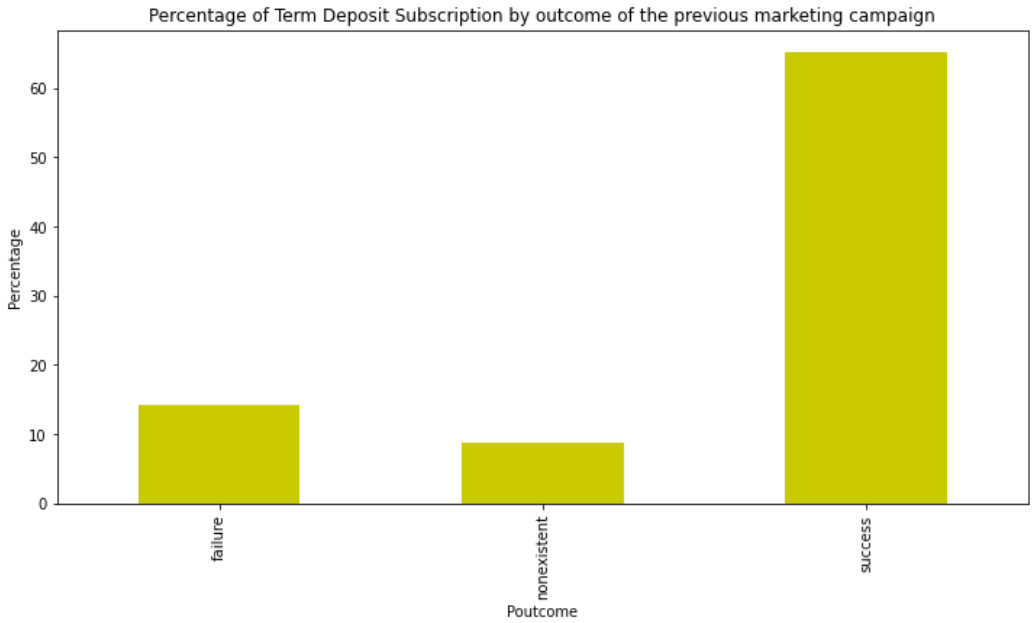
percent of the older people aged 66 and above have subscribed to the plan and the bank needs to do better in young people.



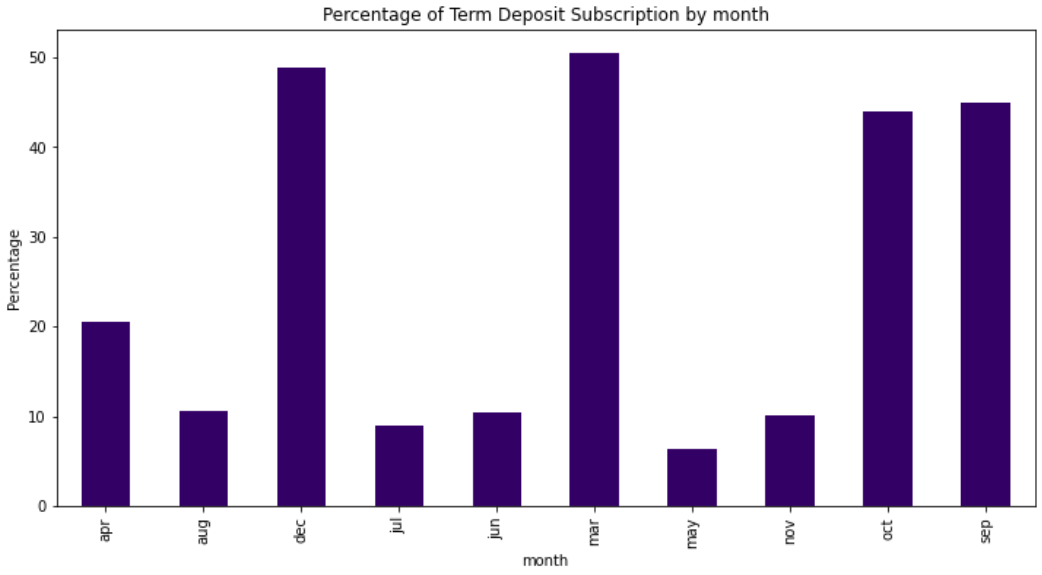
A trend can be found in the way the clients are contacted. Cellular ways seem to provide a better outcome compared to telephone. It can be due to increased usage of mobile phone compared to telephone along with the portability.



From our analysis we have also seen that existing customers who have chosen a previous plan are the major subscribers of the current plan. So, it is evident that the bank was successful in customer retention to a large extent.

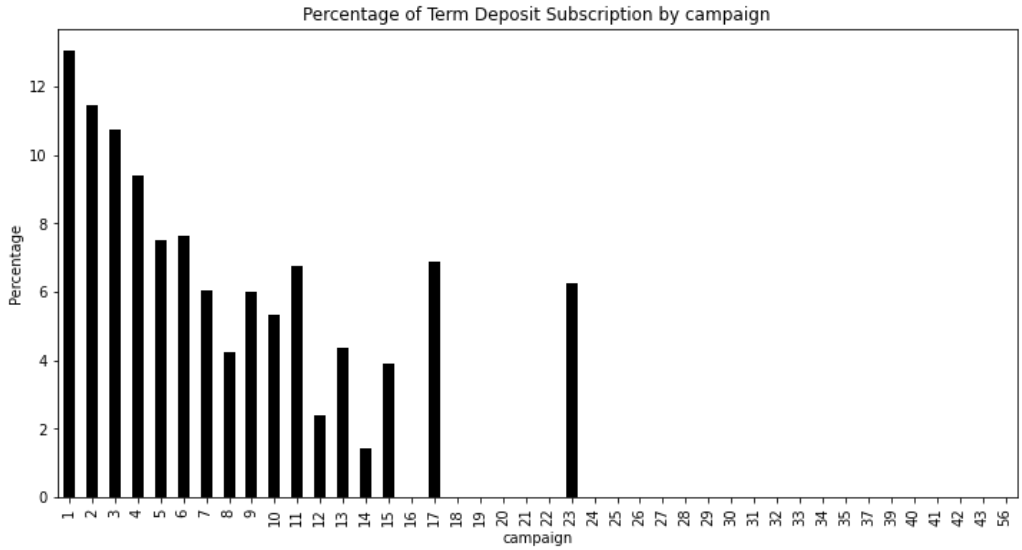


March, September, October, and December months contribute to the major portion of deposit subscriptions. This could be due to some underlying reasons like start of financial year, end of quarter or end of year which should be analyzed further to take wise decisions.

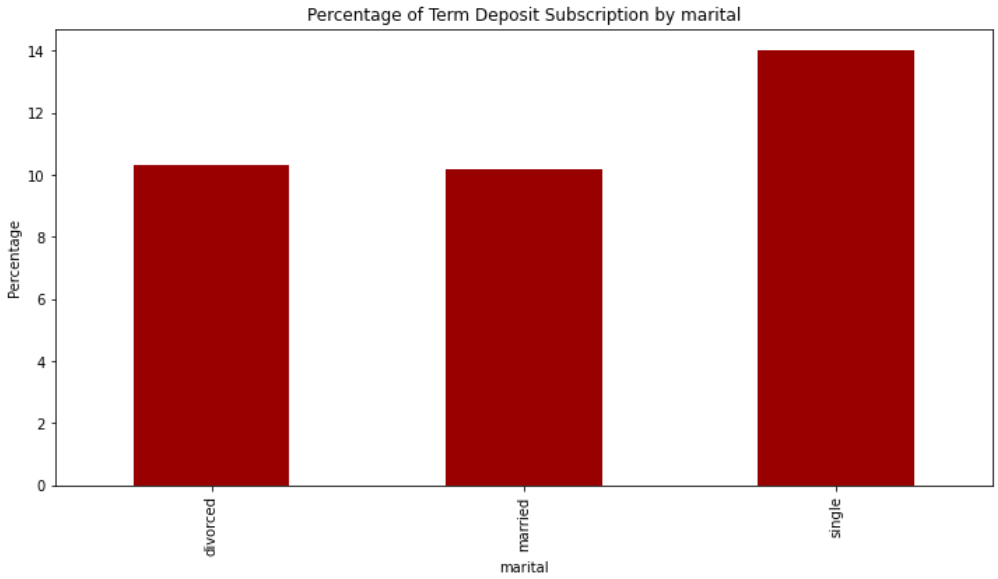


Most of the customers are campaigned only once and yet they are major subscribers of the term deposit. A trend can be seen that even though

many contacts were made to the clients, there is a decline in the percentage of people subscribing to the deposit.

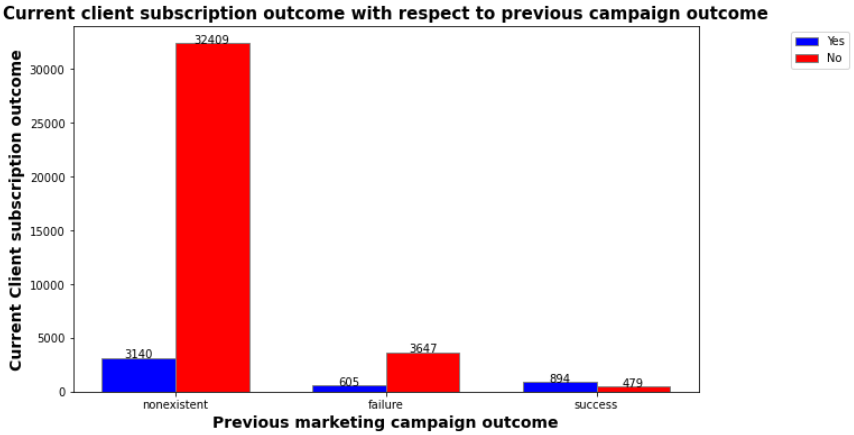


A large percentage of the single clients are subscribing to the term deposit compared to married and widowed. The bank should be exploring ways to attract married and widowed customers and also do more campaigns on single clients.

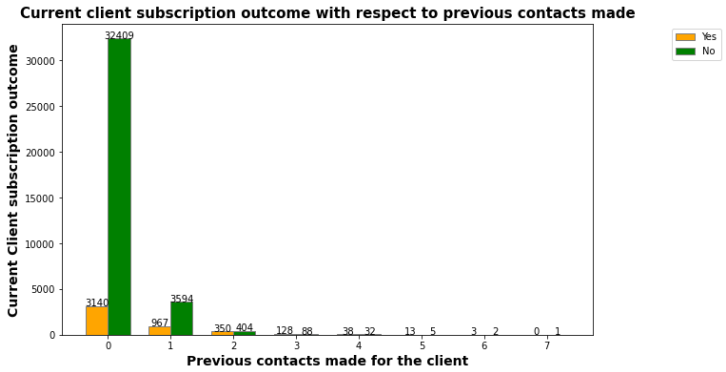


In our analysis, we have observed that many of the clients who have obliged to the previous campaign seem to have subscribed to the current

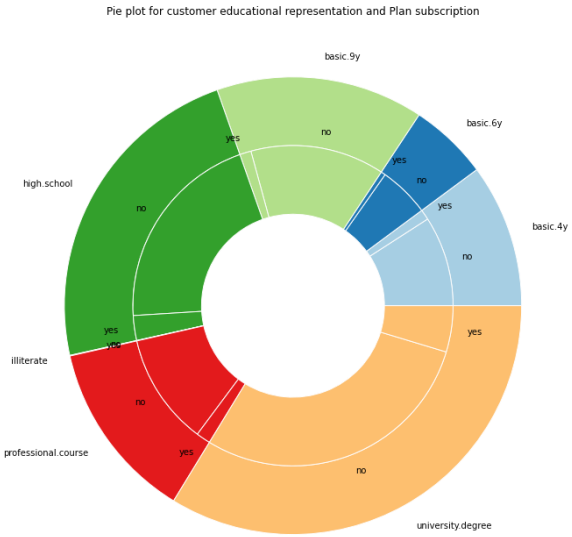
plan as well. Most of the customers who were not part of the previous campaign did not subscribe to the term deposit. Customers who have a failed outcome despite a campaign previously seem to stick to their principles and many did not subscribe to the term deposit on this campaign as well. The Bank needs to do better to attract new customers.



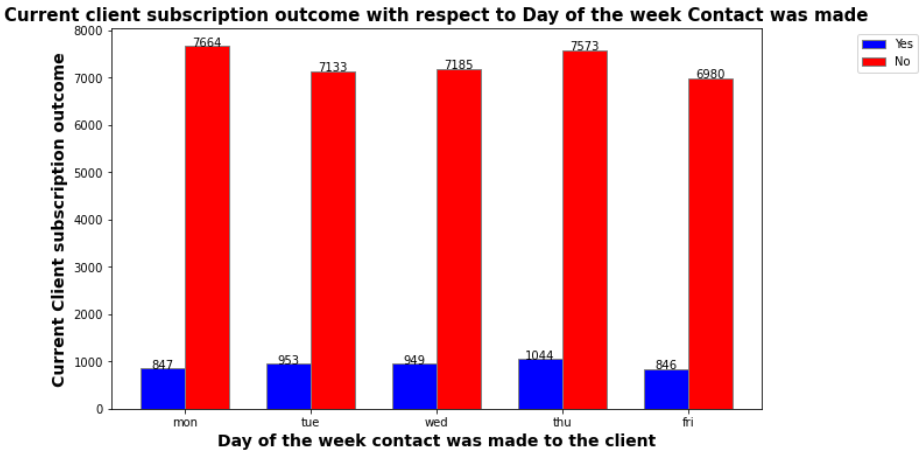
It is clear from the chart that 0 to 7 contacts were made with the client previously. Most of the clients were never contacted previously and most of those clients did not subscribe to the plan even now. One previous contact did not help greatly either. Two contacts made previously started to make a difference where nearly half of them subscribed to term deposit now. Similar is the case with 3 contacts made in which more than half of them made the subscription now. 4 or more calls made previously are a lot lower in number, but still half of the clients subscribed to the current plan. This shows that the company needs to make more calls to attract customers.



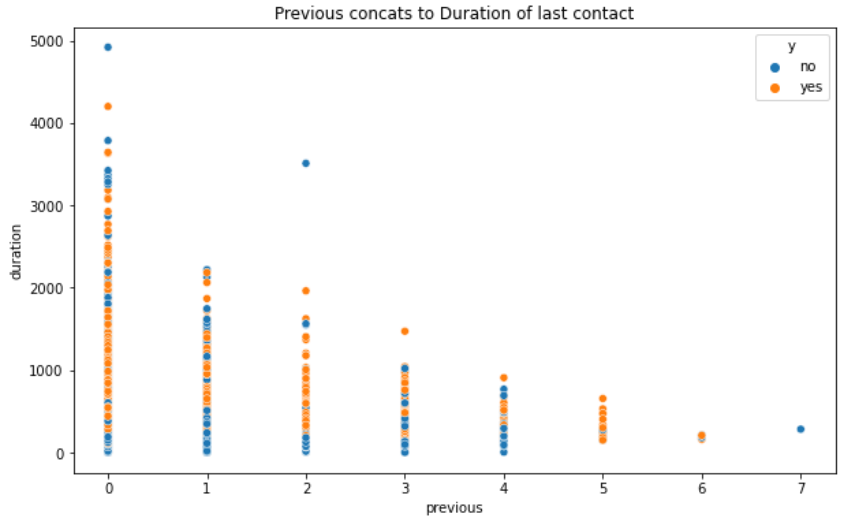
There are 7 categories of educational qualification accomplished by the customers. It includes University degree, Professional course, basic 4 year, basic 9-year, high school and no qualification or illiterate. Out of this there are more people with a university degree. Interestingly most of these people are not subscribing to the plan. Most of the High school qualified clients too are not subscribing to the plan. This trend can be seen among all the classes, including professional degree and basic qualified customers. Overall, the pattern suggests that only 1/5 of the customers are subscribing to the plan irrespective of the educational qualification.



There clearly seemed to a very little effect of the day when the contact was made with the client in determining if the customer would subscribe to the deposit plan. Interestingly Tuesday, Wednesday and Thursday have a little better chance of customers subscribing than Monday and Friday. It could be that customers are a little busy on Mondays and more relaxed on Fridays to consider for a term deposit subscription. However, the variations are marginal.

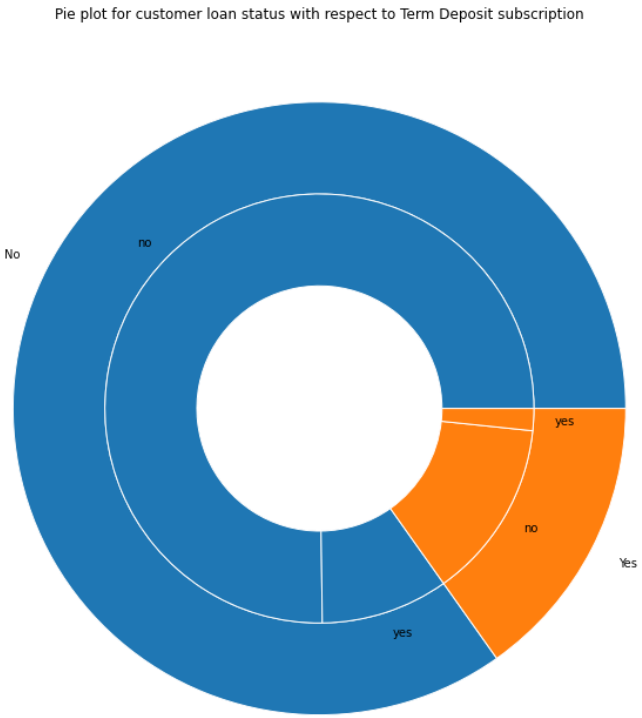


With more previous contacts made with the customer, the duration of the call is greatly reduced. Clients with no previous calls are extending the call duration to until 4000 seconds and most of the customer lot falls in this place. However, customers with more previous calls are subscribing to the plan a lot better even though the number is less.



After mode imputation of unknown values, we see that most of the customers do not have a loan under their name. The outer pie chart represents the loan status and inner pie chart shows the corresponding plan subscription outcome. Loan status, however, is not making any difference because, most of the customers without loan too have not

subscribed to the plan. A few lot of customers who have a loan subscribed to the plan. In a larger view, the ratio between customers who subscribed and customers who did not subscribe with respect to having or not having a loan remains to be the same.



Final Recommendations

- Students and retired individuals are greatly subscribing to the deposit, bank should focus to improve on customers with other employment types as well.
- Elder people are greatly subscribing to the deposit and the bank should invest to attract young people.
- Cellular way of contact is getting more attention from public, and telephone should be targeted only to selected individuals.
- The bank is successful to some extent in retaining customers but should focus on bringing in new customers.

- Months like March, October, September, and December are attracting more customers. The bank should do more campaigns in these months.
- More campaigning and more contacts did not greatly help. The bank should focus on making a better first impression.
- More singles are subscribing to the plan and the bank needs to attract the other set of customers like divorced and married.
- More than half of the customers who have subscribed through a previous campaign have subscribed to the current plan too. The bank needs to focus more on these people.
- Most of the clients contacted had a university degree. The bank should provide more incentives to attract these customers.
- There is a slight edge to attract customers on Tuesday, Wednesday, and Thursday. Therefore, the bank might need to focus more on core weekdays.
- Customers are spending less time on call when contacted more than once. The bank should try to maximize the possibilities in subscribing to the deposit on the first call.
- The bank should not discriminate against clients having a loan because the data clearly shows no difference in the possibilities for people with or without a loan.

Model Selection and Building

We have built models on our dataset with and without using duration column. We have selected decision tree as our base model and a set of models each for every family of Machine learning algorithms. That is Logistic regression in Linear model, Naïve Bayes in classification, KNN in non-linear model, Random Forest in ensemble model, and Adaptive Boost in Boosting.

Without duration column the results are as follows:

Base Model: Decision Tree

Accuracy: 83.69%

Confusion Matrix:
```
[[6575  706]
 [ 637  317]]
```

Classification Model: Naïve Bayes

Accuracy: 21.53 %

Confusion Matrix:
```
[[ 857 6424]
 [  38  916]]
```

Linear Model: Logistic Regression

Accuracy: 89.65 %

Confusion Matrix:
```
[[7201   80]
 [ 772  182]]
```

Non-Linear Model: KNN

Accuracy: 89.11 %

Confusion Matrix:
```
[[7079  202]
 [ 694  260]]
```

Ensemble Model: Random Forest

Accuracy: 89.15 %

Confusion Matrix:
```
[[7064  217]
 [ 676  278]]
```

Boosting Model: ADABoost

Accuracy: 89.39 %

Confusion Matrix:
```
[[7106  175]
 [ 698  256]]
```

Except for naïve bayes, all other models are performing a lot better. Logistic regression has the highest accuracy compared to all other models. Base model i.e. decision tree struggles to some extent.

Adaboost, Random Forest and KNN also does a good job in the prediction.

Considering Duration column in the model, the results are as below:

Base Model: Decision Tree

Accuracy: 88.19%

Confusion Matrix:
```
[[6765  516]
[ 456  498]]
```

Classification Model: Naïve Bayes

Accuracy: 37.47 %

Confusion Matrix:
```
[[2142 5139]
[  10  944]]
```

Linear Model: Logistic Regression

Accuracy: 90.43 %

Confusion Matrix:
```
[[7068  213]
[ 575  379]]
```

Non-Linear Model: KNN

Accuracy: 89.87 %

Confusion Matrix:
```
[[7028  253]
[ 581  373]]
```

Ensemble Model: Random Forest

Accuracy: 91.34 %

Confusion Matrix:
```
[[7042  239]
[ 474  480]]
```

Boosting Model: ADABoost

Accuracy: 91.07 %

Confusion Matrix:
```
[[6997  284]
[ 451  503]]
```

Random forest has the highest accuracy compared to all other models when duration column is included in the features. There is a considerable increase in the accuracy of the decision tree. Naïve Bayes still performs poorly with accuracy of 37%. KNN, logistic regression and boosting models has done good job as before. We can see an overall increase in accuracy in all the models after including the duration column.

Final Recommendation

Classification model naïve bayes is performing poorly and should be avoided. Decision tree is doing a good job to some extent however there are models like logistic regression and random forest that are performing even better. KNN does not seem to have greater impact after including duration column. Random forest, ADAboost and logistic regression have considerable impact after including duration column. With higher accuracy, we would choose Random Forest as our final model due to its consistency and performance irrespective of including or removing the duration column.